Contextual Policy Transfer in Reinforcement Learning Domains via Deep Mixtures-of-Experts (Supplementary Material)

Michael Gimelfarb* **Scott Sanner*** **Chi-Guhn Lee**

Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Canada

Abstract

In this Appendix, we include and provide discussion for additional plots that had to be left out of the main paper due to space limitations. We also describe all hyper-parameters chosen to reproduce the experiments in the main paper.

ADDITIONAL PLOTS

Figure 8 illustrates the test performance on all three transfer learning experiments (Transfer-Maze, Transfer-CartPole and Transfer-SparseLunarLander) using different values of the reuse parameter p_t for the MAPSE algorithm. Figure 9 demonstrates the test performance on all transfer learning experiments using different values of the reuse parameter p_t for the UCB-based policy reuse algorithm. Figure 10 illustrates the test performance on all transfer learning experiments using different learning rates β for the last layer of the option-value network for the option-based context-aware policy reuse algorithm CAPS.

FURTHER IMPLEMENTATION DETAILS

All code was written and executed using Eclipse PyDev running Python 3.7. All neural networks were initialized and trained using Keras with TensorFlow backend (version 1.14), and weights were initialized using the default setting. The Adam optimizer was used to train all neural networks. Experiments were run on an Intel 6700-HQ Quad-Core processor with 8 GB RAM running on the Windows 10 operating system. The hyper-parameter settings used in the experiments are listed in Table 1.

* we had to decrease the learning rate for MARS and reward shaping using a single policy to avoid instability

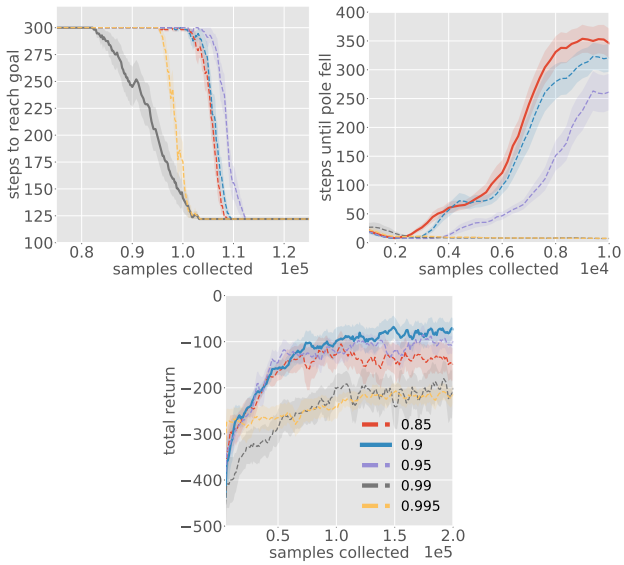

Figure 8: Smoothed mean test performance using the greedy policy for different value of p_t for MAPSE, on Transfer-Maze, Transfer-CartPole, and Transfer-SparseLunarLander (left to right). We ran 20 trials for Transfer-Maze and Transfer-CartPole and 10 trials for Transfer-SparseLunarLander.

** we report the best value found in $\{0.01, 0.001, 0.0001\}$ in the deep learning case
*** $p_t = p^t$ where t is the episode number; we report the best $p \in \{0.85, 0.9, 0.95, 0.99, 0.995\}$

Also, please note that for the imperfect transfer setting, we set $c = 0.001$ for MARS and reduce the learning rate to 0.6, since the algorithm converges much slower for larger values of c or becomes unstable.

*Affiliate to Vector Institute, Toronto, Canada.

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

Parameters				
Name	Description	Transfer-Maze	Transfer-CartPole	Transfer-SparseLunarLander
T	maximum roll-out length	300	500	1000
γ	discount factor	0.95	0.98	0.99
ε_t	exploration probability	0.12	$\max\{0.01, 0.99^t\}$	$\max\{0.01, 0.9925^t\}$
	learning rate of Q-learning*	0.08, 0.8		
	replay buffer capacity		5000	20000
B	batch size		32	64
Q	topology of DQN		4-40-40-4	8-120-100-4
	hidden activation of DQN		ReLU	ReLU
	learning rate of DQN*		0.0002, 0.0005	0.0002, 0.0005
	learning rate for termination function weights**	0.4	0.01	0.0001
	target network update frequency (in batches)		500	100
	L2 penalty of DQN		10^{-6}	10^{-6}
$\hat{f}_i$	topology of dynamics model		8-50-50-4	12-100-100-8
	hidden activation of dynamics model		ReLU	ReLU
	learning rate of dynamics model		0.001	0.001
	L2 penalty of dynamics model		10^{-6}	10^{-6}
ν_i	Gaussian kernel precision		5×10^5	5×10^5
$\mathbf{a}$	topology of mixture model	58-30-30-4	4-30-30-3	8-30-30-3
	hidden activation of mixture	ReLU	ReLU	ReLU
λ	learning rate of mixture	0.001	0.001	0.001
	training epochs/batch for mixture	4	3	1
c	PBRS scaling factor	1.0	2.0	20.0
p_t	probability of following source policies***	0.99^t (MAPSE), 0.85^t (UCB)	0.85^t (MAPSE), 0.85^t (UCB)	0.9^t (MAPSE), 0.95^t (UCB)

Table 1: Hyper-parameter settings.

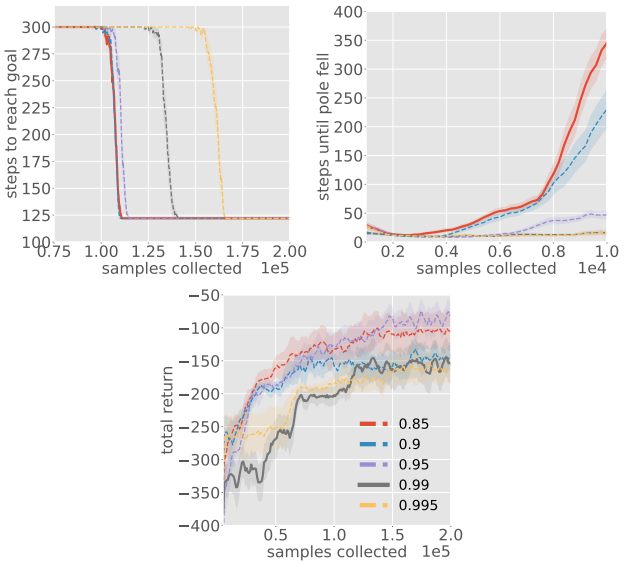

Figure 9: Smoothed mean test performance using the greedy policy for different value of p_t for UCB, on Transfer-Maze, Transfer-CartPole, and Transfer-SparseLunarLander (left to right). We ran 20 trials for Transfer-Maze and Transfer-CartPole and 10 trials for Transfer-SparseLunarLander.

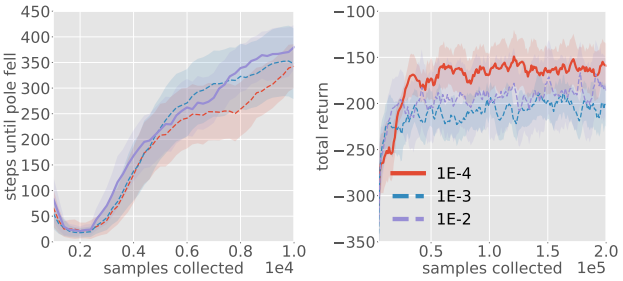

Figure 10: Smoothed mean test performance using the greedy policy for different value of termination learning rate for CAPS. From left to right: (1) number of steps balanced on Transfer-CartPole, and (2) total return on Transfer-SparseLunarLander. We ran 20 trials for Transfer-CartPole and 10 trials for Transfer-SparseLunarLander.